CSIT 5300 - Spring 2020 - HKUST
Assignment 1

Deadline:  
Tuesday, May 19, 2020, 23:59
Submission: Submit a soft copy of your solution to Canvas

<table>
<thead>
<tr>
<th>Student Name</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Student ID</td>
<td></td>
</tr>
<tr>
<td>Email Address</td>
<td></td>
</tr>
<tr>
<td>Section</td>
<td></td>
</tr>
</tbody>
</table>

For T.A. use only

<table>
<thead>
<tr>
<th>Problem</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>/ 70</td>
</tr>
<tr>
<td>2</td>
<td>/ 10</td>
</tr>
<tr>
<td>3</td>
<td>/ 10</td>
</tr>
<tr>
<td>4</td>
<td>/ 10</td>
</tr>
<tr>
<td>Total</td>
<td>/ 100</td>
</tr>
</tbody>
</table>
Textbooks on Databases (or any other Computer Science area) typically become obsolete as technology advances. The most up-to-date systems and algorithms are published in conference and journal articles, which become available much faster than the new textbook versions. Big companies often watch closely the conference and journal publications, in order to integrate the new technologies to their products as soon as possible, or devise new and better systems before anybody else (claiming patents to safeguard their technology).

The purpose of this assignment is to test your ability in comprehending a research article. You are required to select exactly one research article, out of a set of 5 available articles, and solve 4 problems outlined for this article. The first problem always asks you to write a brief summary of the article, and is assigned the biggest grade contribution (70%). The other three problems are simple questions on the content of the article. Pages 3-7 below describe these 5 articles and their problems, respectively. Note that you may have to be on the UST campus to be able to download the PDF files of these articles from the provided links.
**Topic #1:** Keyword Search in Relational Databases

**Article:** S. Agrawal, S. Chaudhuri, G. Das. DBXplorer: A System for Keyword-Based Search over Relational. In *ICDE*, 2002.

**Link:** [http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=994693&tag=1](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=994693&tag=1)

**Problem 1** [70 points]
Give a brief summary of the paper. The summary should provide a concise idea of the problem and the solution, and it should not exceed the 150 words. Please try to limit each point that you want to mention to the essence of the idea without giving many details.

**Problem 2** [10 points]
Give a short description of the *publish* and *search* step.

**Problem 3** [10 points]
Given a keyword-based query, what is DBXplorer supposed to return?

**Problem 4** [10 points]
Describe the *row* and *column* level granularities briefly. Assuming that there is no index on any attribute, which level granularity is better? Justify your answer by taking into account the space and time requirements, the ease of maintenance and the keyword search performance of each approach.
**Topic #2:** Data mining in Relational Tables


**Problem 1** [70 points]
Give a brief summary of the paper. The summary should provide a concise idea of the problem and the solution, and it should not exceed the 150 words. Please try to limit each point that you want to mention to the essence of the idea without giving many details.

**Problem 2** [10 points]
Explain briefly what *minimum support* and *minimum confidence* is.

**Problem 3** [10 points]
Consider the database $D$ consisting of 9 transactions:
\[
\{\{1,2,5\}, \{2,4\}, \{2,3\}, \{1,2,4\}, \{1,3\}, \{2,3\}, \{1,3\}, \{1,2,3,5\}, \{1,2,3\}\}.
\]
Suppose that the minimum support is 20%. Show the steps of the algorithm for computing the itemsets $L_i$, where $1 \leq i \leq 4$.

**Problem 4** [10 points]
Let the minimum confidence be 70%. Identify the valid association rules of the database $D$ of Problem 3. You can use the result you found in Problem 3.
Topic #3: Data Streams


Problem 1 [70 points]
Give a brief summary of the paper. The summary should provide a concise idea of the problem and the solution, and it should not exceed the 150 words. Please try to limit each point that you want to mention to the essence of the idea without giving many details.

Problem 2 [10 points]
Give a short description of the three classes of algorithms that solve the problem.

Problem 3 [10 points]
Enumerate the pros and cons of each algorithm class.

Problem 4 [10 points]
According to the experimental evaluation, which is the most efficient algorithm of each class (in terms of number of updates per ms and precision)? Give a brief explanation.
**Topic #4: XML**


**Problem 1** [70 points]
Give a brief summary of the paper. The summary should provide a concise idea of the problem and the solution, and it should not exceed the 150 words. Please try to limit each point that you want to mention to the essence of the idea without giving many details.

**Problem 2** [10 points]
What is XML and why it is useful? Which is the main difference of XML and Relational Databases?

**Problem 3** [10 points]
What is the fragmentation problem? Give an example using a short DTD and a simple query.

**Problem 4** [10 points]
Give the pros and cons of the Shared and Hybrid Inlining techniques.
Problem 1 [70 points]
Give a brief summary of the paper. The summary should provide a concise idea of the problem and the solution, and it should not exceed the 150 words. Please try to limit each point that you want to mention to the essence of the idea without giving many details.

Problem 2 [10 points]
Give a short description of the data model.

Problem 3 [10 points]
Describe the responsibilities of the master and the tablet server.

Problem 4 [10 points]
How does a client know which tablet server contains the data he/she needs?